Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: https://orca.cardiff.ac.uk/id/eprint/92137/

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Al-Wakeel, Ali ORCID: https://orcid.org/0000-0003-4970-7309, Wu, Jianzhong ORCID: https://orcid.org/0000-0001-7928-3602 and Jenkins, Nicholas ORCID: https://orcid.org/0000-0003-3082-6260 2017. k-means based load estimation of domestic smart meter measurements. Applied Energy 194 , pp. 333-342. 10.1016/j.apenergy.2016.06.046 file

Publishers page: http://dx.doi.org/10.1016/j.apenergy.2016.06.046 <http://dx.doi.org/10.1016/j.apenergy.2016.06.046>

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

information services
gwasanaethau gwybodaeth

Applied Energy xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

k-means based load estimation of domestic smart meter measurements ☆

Ali Al-Wakeel, Jianzhong Wu *, Nick Jenkins

School of Engineering, Cardiff University, Cardiff CF24 3AA, United Kingdom

HIGHLIGHTS

- A load estimation algorithm based on *k*-means cluster analysis was developed.
- Canberra, Manhattan, Euclidean, and Pearson correlation distances were investigated.
- Daily and segmented load profiles of aggregated smart meters were used.
- Canberra distance outperforms the other distance functions.
- High accuracy estimates were obtained with cluster centres between 16 and 24 h.

ARTICLE INFO

Article history:
Received 16 March 2016
Received in revised form 22 May 2016
Accepted 12 June 2016
Available online xxxx

Keywords:
Cluster analysis
k-means
Smart meter measurements
Load estimation

ABSTRACT

A load estimation algorithm based on *k*-means cluster analysis was developed. The algorithm applies cluster centres – of previously clustered load profiles – and distance functions to estimate missing and future measurements. Canberra, Manhattan, Euclidean, and Pearson correlation distances were investigated. Several case studies were implemented using daily and segmented load profiles of aggregated smart meters. Segmented profiles cover a time window that is less than or equal to 24 h. Simulation results show that Canberra distance outperforms the other distance functions. Results also show that the segmented cluster centres produce more accurate load estimates than daily cluster centres. Higher accuracy estimates were obtained with cluster centres in the range of 16–24 h. The developed load estimation algorithm can be integrated with state estimation or other network operational tools to enable better monitoring and control of distribution networks.
© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

The installation of smart meters is usually considered as the starting point in the implementation of Smart Grids [1]. Smart meters employ advanced metering, control, data storage, and communication technologies to offer a range of functions. Nearly 53 million gas and electricity smart meters will be installed in all domestic and small non-domestic premises in Great Britain by the end of 2020 [2,3].

The deployment of smart meters provides benefits to the end consumers (domestic and non-domestic), energy suppliers, and network operators by providing near real-time consumption information to the consumers that will help them to manage their energy use, save money, and reduce greenhouse gas emissions [3]. At the same time, smart meters will benefit distribution network planning and operation, and demand management. In this regard, the smart metering data will enable more accurate demand forecasts, allow improved asset utilisation in distribution networks, locate outages and shorten supply restoration times, and reduce the operational and maintenance costs of the networks [4,5].

Smart meters provide volumes of data ranging from several hundreds of gigabytes to tens of petabytes (or exabytes) for the energy suppliers and network operators to exploit [6–8]. The data volume will vary according to the number of installed smart meters, the number of received smart meter messages, the message size (in bytes per message), and the frequency of recording the measurements – e.g., every 15 or 30 min.

Great Britain's smart metering system faces significant technical and operational challenges. The technical challenges include intermittent communication networks (both mobile and radio frequency); the lack of sufficient signal strength; the shortage of tools

☆ The short version of the paper was presented at CUE2015 on November 15–17, Fuzhou, China. This paper is a substantial extension of the short version.

* Corresponding author.
E-mail address: wuj5@cardiff.ac.uk (J. Wu).

http://dx.doi.org/10.1016/j.apenergy.2016.06.046
0306-2619/© 2016 The Authors. Published by Elsevier Ltd.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Please cite this article in press as: Al-Wakeel A et al. *k*-means based load estimation of domestic smart meter measurements. Appl Energy (2016), http://dx.doi.org/10.1016/j.apenergy.2016.06.046

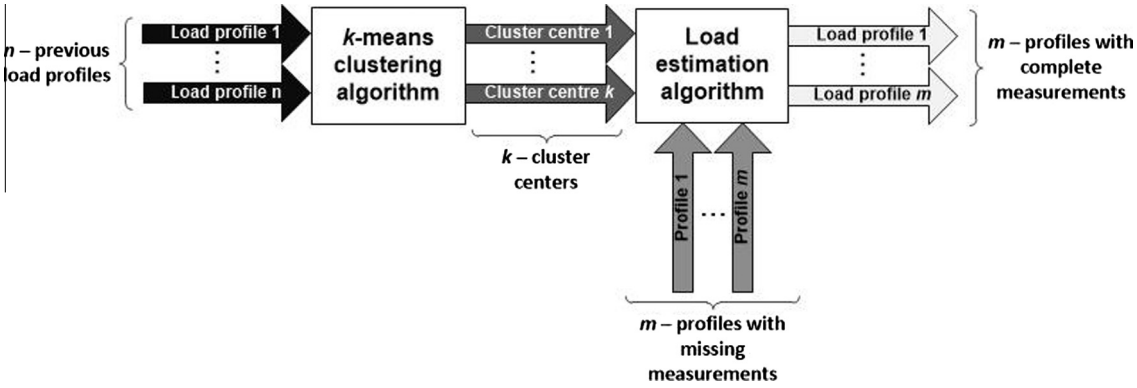

Fig. 1. Structure of the developed load estimator.

to detect mobile network failure; and the indoor/outdoor placement of meters. Examples of the operational challenges include planned or unplanned maintenance of the system, software and hardware faults or malfunction of the smart meters, and customers unwilling to communicate their energy consumption data. These challenges make smart meter measurements susceptible to time delays or even temporary loss when requested by the energy suppliers or network operators [9–11]. It is therefore necessary to develop a load estimation algorithm to replace missing and estimate future smart meter measurements. In this sense, load estimation analyses the past measurements and extracts practical information – e.g., the load profiles of typical customers – to estimate the missing measurements.

Statistical, engineering, and time-series methods [12,13] have been reported to analyse and extract the required information from the load profiles of customers. Additionally, statistical, time-series and artificial intelligence (AI) methods have been applied to estimate and forecast the load in power networks [14–17]. However, these methods can be costly and complex to implement and validate when large volumes of consumption measurements become available. One efficient approach to extract the necessary information from smart meter measurements is the employment of data mining techniques. Cluster analysis is one type of data mining techniques [18,19]. Several clustering methods have been reported to group the load profiles of different types of customers in distribution networks. In the previous research [12,20,21], cluster analysis methods were applied to develop the typical daily load profiles (TDLPs) of different types of customers for pricing and settlement purposes. Nonetheless, the application of cluster analysis methods to solve the load estimation problem has been limited [22,23].

This paper proposes a load estimation algorithm that was developed using the *k*-means cluster analysis method. The load estimation algorithm is easy to implement and requires no prior knowledge of any variables other than historical half-hourly power consumption measurements. The algorithm provides a good compromise between the quality of the solution and the computational complexity and therefore can be used to estimate missing and future measurements of aggregated smart meters at the medium voltage (MV) level. The accuracy of the estimated measurements makes them applicable for a variety of network operational tools so as to enable better monitoring and control of distribution networks.

Additionally, the developed load estimation algorithm overcomes the key drawbacks of *k*-means cluster analysis method, such as the impact of initial selection of cluster centres and the necessity to predefine the required number of clusters. Furthermore, this work investigates the applicability of daily and segmented cluster centres for load estimation. To the best of authors' knowledge, clustering the segmented load profiles and the application of segmented cluster centres for load estimation has not been reported in previous research.

Fig. 1 shows the high-level structure of the developed load estimation algorithm. The *k*-means method [24,25] was applied to group similar load profiles and produce a number of cluster centres. These centres were used to estimate the smart meter measurements using different distance functions.

2. Cluster analysis methods

Clustering is the grouping of load profiles into a number of clusters such that profiles within the same cluster are similar to each other. At the same time, load profiles that are assigned to different clusters are as dissimilar as possible. In this manner, the profiles are clustered based on the principle of "maximising the intra cluster similarity and minimising the inter cluster similarity[1]".

Clustering implies that the number of output clusters is less than or equal to the number of input load profiles. Applications of clustering include classification, pattern recognition, and clustering based estimation. Large numbers of cluster analysis methods have been developed [18,24]. Cluster analysis methods are broadly categorised into hierarchical and partitional clustering methods.

Hierarchical methods [26] group a given dataset of load profiles into the required number of clusters through a series of nested partitions. This results in a hierarchy of partitions leading to the final cluster(s).

Partitional methods on the other hand represent each cluster by a centre, which is a summary description of all load profiles contained within the cluster. Partitional methods aim to group load profiles into a number of clusters by optimising an objective function. The distance between the profiles and cluster centre is the objective function that is minimised [27]. In partitional clustering, the required number of clusters must be predefined or known in advance.

The *k*-means [24] is a classic partitional cluster analysis method and has been reported to group the load profiles of customers in power networks. Table 1 illustrates the merits and drawbacks of this method; and lists significant research papers that report the application of the *k*-means method for load profile characterization in power networks.

3. Data structure

Domestic load profiles based on smart meter measurements were used to investigate the performance of the developed load

[1] This is the same as maximising both the intra cluster similarity and the inter cluster dissimilarity.

Please cite this article in press as: Al-Wakeel A et al. *k*-means based load estimation of domestic smart meter measurements. Appl Energy (2016), http://dx.doi.org/10.1016/j.apenergy.2016.06.046

Table 1
Summary of the *k*-means cluster analysis method.

Clustering method	Advantages	Disadvantages	Refs.
k-means	• Simple • Efficient • Scalable • Ability to handle big data • Linear complexity with the size of dataset	• Sensitive to the initial selection of cluster centres • Number of clusters must be defined in advance • Sensitive to noise and outliers • Provides local (not global) optimum solution	[12,20,21,24,28]

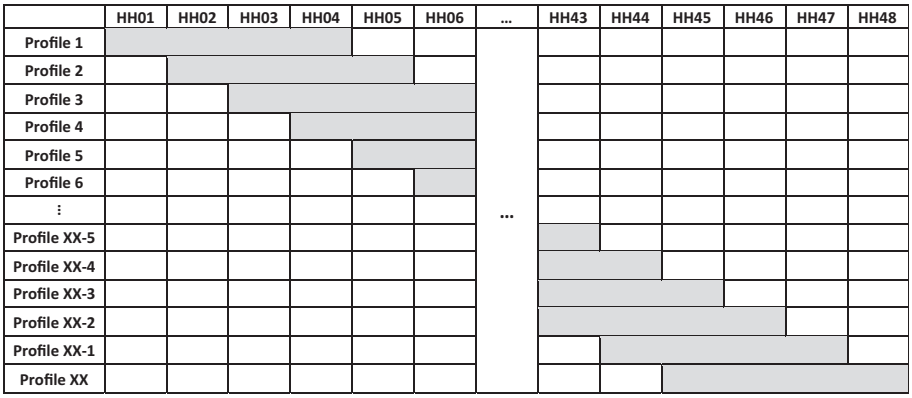

Fig. 2. Segmented load profiles.

estimation algorithm. The load profiles were obtained from the Irish Smart Metering Customer Behaviour Trials (CBT) which were accessed via the Irish Social Science Data Archive [29]. The Irish CBT is one of the largest and most statistically robust smart metering trials. These trials were implemented to investigate the impact of smart meter technology upon power consumption behaviour for different types of customers, and to identify a "Tipping point [2]". Customers' behaviour in terms of peak demand and overall electricity consumption was analysed combining smart meter technology with time-of-use tariffs and demand side management stimuli. The trials were carried out in the course of 18 months from 1st July 2009 until 31st December 2010. More than 4200 domestic customers and 485 small-and-medium enterprises (SMEs) were covered by these trials. Smart meters – that were installed at the customers' premises – recorded the consumption data. For an individual customer (smart meter), 48 half-hourly average active power consumption measurements represent any daily load profile. The first measurement – that was recorded at hour 00:30 – is the average power consumed between hours 00:00:00 and 00:29:59, whereas the last measurement – that was recorded at hour 00:00 – is the average power consumption between hours 23:30:00 and 23:59:59.

Two weeks of smart meter measurements of 100 randomly selected domestic customers were used in this study. The measurements collected between 20th July and the end of 26th July 2009 were applied to train the *k*-means clustering method. A test set of measurements – over the period from 27th July until the end of 3rd August 2009 – was used for load estimation.

This study investigates the estimation of aggregated smart meter measurements. An aggregated daily load profile was created by summing the measurements of the 100 smart meters at each half hour time step. Eq. (1) illustrates the aggregation of smart meter measurements.

$$LP_{agg.daily} = \left\{ \left(\sum_{i=1}^{100} lp_i(t) \right)_{t=1}, \left(\sum_{i=1}^{100} lp_i(t) \right)_{t=2}, \ldots, \left(\sum_{i=1}^{100} lp_i(t) \right)_{t=48} \right\} \quad (1)$$

Daily and segmented load profiles were initially clustered. A daily load profile consists of 48 half-hourly measurements; whereas a segmented load profile extends over a time window that is less than or equal to 24 h. Time windows in the range of 2–24 h were used – on a rolling basis – to create the segmented load profiles. For any segmentation time window (r) in hours, provided that the segmented profiles are rolled one half hourly step at a time, then the number of segmented profiles is determined according to Eq. (2) [30]

$$Number\ of\ segmented\ profiles = (n \times T) - 2r + 1 \quad (2)$$

given that n is the number of the daily load profiles, and T is the number of half-hourly measurements per daily load profile. Daily and segmented clusters centres were separately applied to estimate missing and future smart meter measurements. Fig. 2 illustrates the concept of segmented load profiles.

4. Load estimation methodology

4.1. k-means cluster analysis

The *k*-means method iteratively groups n load profiles – each comprised of T half-hourly measurements – into k clusters, by minimising the intra-cluster sum of squared distances between the load profiles and cluster centres. Eq. (3) illustrates the objective function of the *k*-means method [18,24].

$$J = \sum_{j=1}^{k} \sum_{i=1, i \in j}^{n} \|\mathbf{LP_i} - \mathbf{CC}_j\|^2 \quad (3)$$

$\mathbf{LP_i}$ is a vector that represents the i^{th} load profile, $i = 1, 2, 3, \ldots, n$, and $\mathbf{CC_j}$ is vector representing the j^{th} cluster centre, $j = 1, 2, 3, \ldots, k$. The

[2] Tipping point is the point at which a significant change in consumption is stimulated by the price of electricity [11].

Please cite this article in press as: Al-Wakeel A et al. *k*-means based load estimation of domestic smart meter measurements. Appl Energy (2016), http://dx.doi.org/10.1016/j.apenergy.2016.06.046

i^{th} load profile is described as $\mathbf{LP_i} = [lp_i(t), t = 1, 2, 3, \ldots, T]$. Similarly, the j^{th} cluster centre is defined as $\mathbf{CC_j} = [cc_j(t), t = 1, 2, 3, \ldots, T]$. A cluster centre is determined in terms of the average values of all load profiles assigned to this specific cluster, calculated at each half-hourly time step. Eq. (4) defines the centre of a cluster.

$$\mathbf{CC}_j = \left[\left(\frac{\sum_{i=1}^{m} lp_i(t)}{m}\right)_{t=1}, \left(\frac{\sum_{i=1}^{m} lp_i(t)}{m}\right)_{t=2}, \ldots, \left(\frac{\sum_{i=1}^{m} lp_i(t)}{m}\right)_{t=T}\right] \quad (4)$$

The inputs of the *k*-means based clustering algorithm include the training load profiles, the randomisation number (*rnd_m*), and the maximum number of clusters. The *k*-means algorithm – that was developed – defines the randomisation number as a variable to overcome the impact of the initial random selection of cluster centres upon the outputs. For a number of (*k*) clusters, the randomisation number runs the *k*-means method *rnd_m* different times; each time with a different set of initial cluster centres. The best results – those with the smallest intra cluster distances – are produced as outputs.

At each iteration of the *k*-means, the Average Euclidean distance (*AED*) is calculated between the input load profiles and their cluster centres according to Eq. (5). As a result, each load profile is assigned to the cluster that has the nearest centre.

$$AED(\mathbf{LP_i}, \mathbf{CC_j}) = \sqrt{\frac{\sum_{t=1}^{T} (lp_i(t) - cc_j(t))^2}{T}} \quad (5)$$

Eq. (6) defines the mean *AED* that was used as a criterion to determine the required number of clusters. In this study, the number of clusters is incremented until the mean *AED* falls below one percent of the average active power consumption of the training period.

$$Mean\ AED = \underset{n}{\text{mean}}(AED(\mathbf{LP}, \mathbf{CC})) = \frac{\sum_{i=1}^{n} AED(\mathbf{LP_{i,i\in j}}, \mathbf{CC_j})}{n} \quad (6)$$

The outputs of the clustering algorithm include the number of clusters, cluster centres, and load profiles assigned to their respective clusters.

Fig. 3 shows a flow chart of the proposed *k*-means cluster analysis algorithm. Pycluster [31], an open source cluster analysis software was used to develop the clustering module in Python 2.7.

4.2. Load estimation

The developed load estimator applies the cluster centres – the outputs of the *k*-means cluster analysis algorithm – to estimate any missing measurements in the test period. Distance functions were used to link test profiles – with missing measurements – to the nearest training cluster centre. For each day of the test period, 24 scenarios of lost measurement were simulated using a brute-force approach. The scenarios consider different durations of lost measurements (T_{loss}) from 1 to 24 consecutive hours. The measurements were estimated iteratively, i.e., only one half-hourly measurement was estimated at a time.

The root-mean-square error (RMSE) and mean absolute percentage error (MAPE) were applied to quantify the errors between the estimated and actual measurements of the test period. Eq. (7) illustrates the RMSE of load estimates, whereas Eq. (8) defines the MAPE of the estimated load.

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{T_{loss}} (lp_{act,n}(t) - lp_{est,n}(t))^2}{T_{loss}}} \quad (7)$$

$$\text{MAPE} = \frac{1}{T_{loss}} \sum_{t=1}^{T_{loss}} \left| \frac{lp_{act,n}(t) - lp_{est,n}(t)}{lp_{act,n}(t)} \right| \quad (8)$$

$lp_{act,n}(t)$ is the actual half-hourly measurement, $lp_{est,n}(t)$ is the estimated half-hourly measurement, t is the half-hour index, T_{loss} is the duration of lost measurements in hours, and n is the sample index.

4.2.1. Estimation using daily cluster centres

Forty-eight half-hourly measurements (the measurement to be estimated plus 47 half-hourly measurements that precede it) were matched to the nearest daily cluster centre. Fig. 4 illustrates this approach.

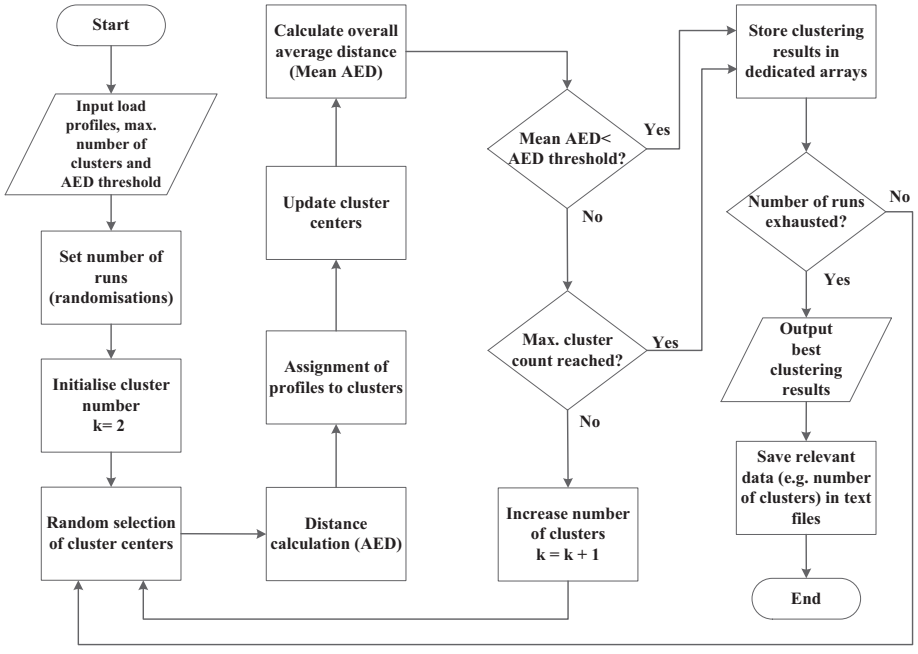

Fig. 3. Developed *k*-means cluster analysis algorithm.

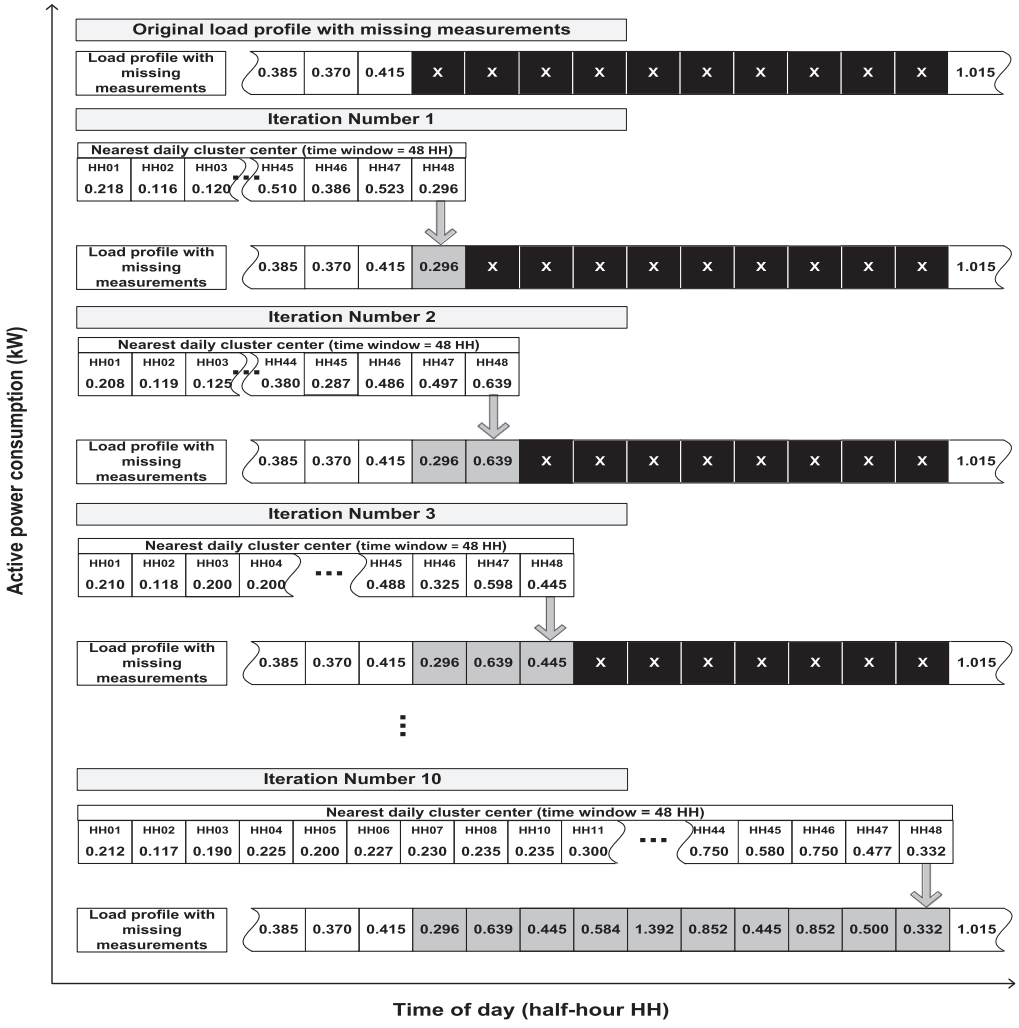

Fig. 4. Load estimation using daily cluster centres.

4.2.2. Estimation using segmented cluster centres

Rather than use a time window of 24 h, $2r$ half-hourly measurements (one measurement to be estimated plus the $2r-1$ half-hourly measurements that precede it) were paired to the nearest segmented cluster centre – whose length is $2r$ half-hourly measurements. Fig. 5 illustrates this approach.

4.2.3. Distance functions

The load estimator investigated the application of four different distance functions to produce the required load estimates. Table 2 illustrates the distance functions that were used.

Euclidean distance (D1) is the most commonly used distance function in engineering and physical sciences. Manhattan (also called city block, taxicab and rectilinear) distance (D2) measures distances on a rectilinear basis. Canberra distance (D3) is considered a special case of Manhattan distance. The only difference between these two measures is that in Canberra distance, the absolute differences between the k^{th} instances of the load profiles are divided by the sum of the absolute values of these instances prior to summing all instances. Measure (D4) is a dissimilarity measure rather than an actual distance metric [24,32,33]. It is derived from the Pearson correlation coefficient applying Eq. (9)

$$d(LP_i, LP_j) = 1 - s(LP_i, LP_j) \tag{9}$$

where $s(LP_i, LP_j)$ is the Pearson correlation coefficient described by Eq. (10),

$$s(LP_i, LP_j) = \frac{\sum_{t=1}^{T}(lp_i(t) - \overline{LP}_i)((lp_j(t) - \overline{LP}_j))}{\left[\sum_{t=1}^{T}(lp_i(t) - \overline{LP}_i)^2 \sum_{t=1}^{T}(lp_j(t) - \overline{LP}_j)^2\right]^{1/2}} \tag{10}$$

5. Results and discussion

The load estimation algorithm was applied to estimate the load measurements of an MV busbar. Figs. 6 and 7 show the actual and estimated load profiles for a weekday and a weekend.

The solid black profile is the actual load profile that was obtained using real aggregated smart meter measurements collected from domestic premises. The solid red profile is the mean of all estimated profiles that were produced by the load estimation algorithm. The dashed red profiles are the mean of the minimum and maximum estimated profiles.

5.1. Impact of the distance function

Simulation results reveal that the application of Canberra distance yields more accurate load estimates than other distance functions. A box-whisker plot [34] of the MAPE and RMSE of the estimated measurements is shown in Fig. 8. Regardless of both the duration of lost measurements and time window of segmented cluster centres, Fig. 8 shows that the application of Canberra and

Please cite this article in press as: Al-Wakeel A et al. *k*-means based load estimation of domestic smart meter measurements. Appl Energy (2016), http://dx.doi.org/10.1016/j.apenergy.2016.06.046

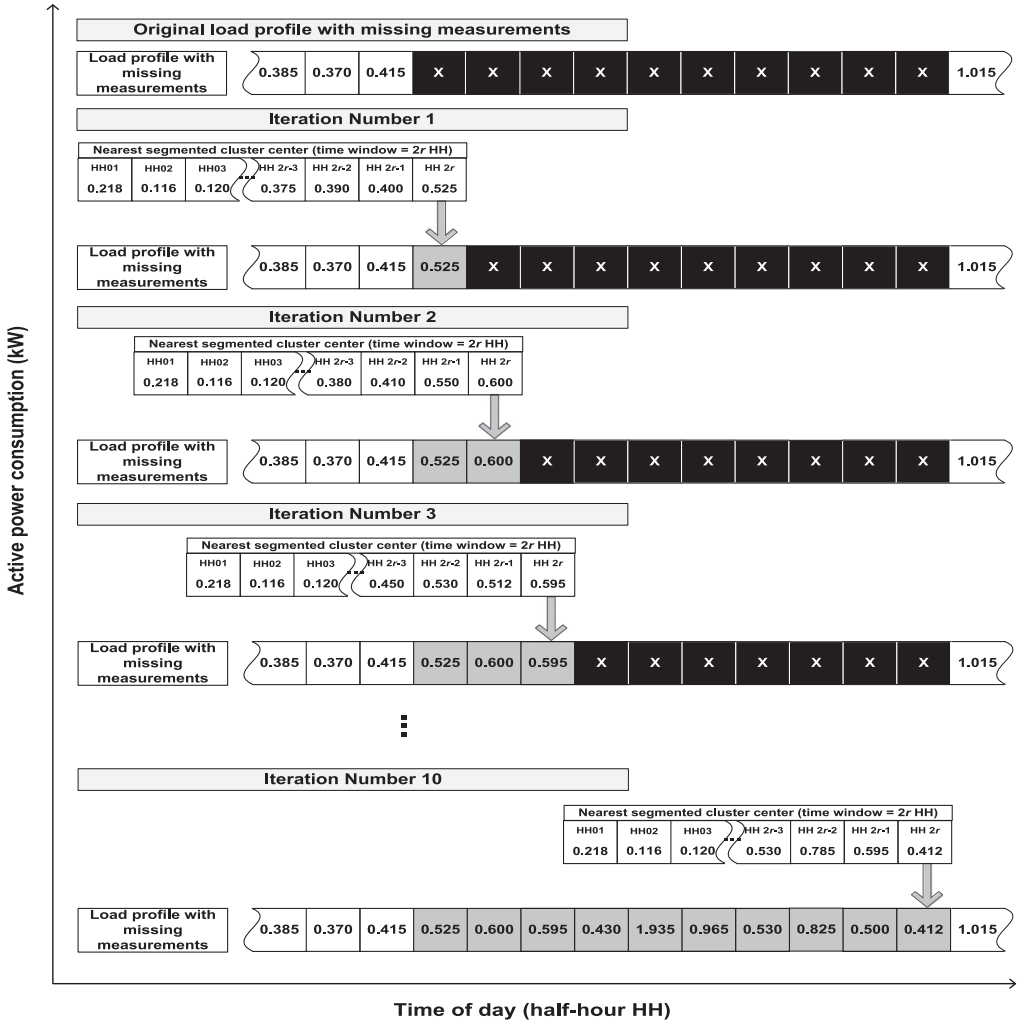

Fig. 5. Load estimation using segmented cluster centres.

Table 2
Summary of the distance functions.

Measure	Formula
D1: Average Euclidean distance	$d(LP_i, LP_j) = \left[\frac{\sum_{t=1}^{T}(lp_i(t)-lp_j(t))^2}{T}\right]^{1/2}$
D2: Average Manhattan (city block) distance	$d(LP_i, LP_j) = \frac{\sum_{t=1}^{T}\lvert lp_i(t)-lp_j(t)\rvert}{T}$
D3: Average Canberra distance	$d(LP_i, LP_j) = \begin{cases} 0 & \text{for } lp_i(t)=lp_j(t)=0 \\ \frac{\sum_{t=1}^{T}\frac{\lvert lp_i(t)-lp_j(t)\rvert}{\lvert lp_i(t)\rvert+\lvert lp_j(t)\rvert}}{T} & \text{for } lp_i(t)\neq 0 \text{ or } lp_j(t)\neq 0 \end{cases}$
D4: Average Pearson correlation distance	$d(LP_i, LP_j) = \frac{1-\frac{\sum_{t=1}^{T}(lp_i(t)-\overline{LP}_i)((lp_j(t)-\overline{LP}_j))}{\left[\sum_{t=1}^{T}(lp_i(t)-\overline{LP}_i)^2\sum_{t=1}^{T}(lp_j(t)-\overline{LP}_j)^2\right]^{1/2}}}{T}$ where $\overline{LP}_i$ is the average value of the ith load profile and $\overline{LP}_j$ is the mean values of the jth load profile

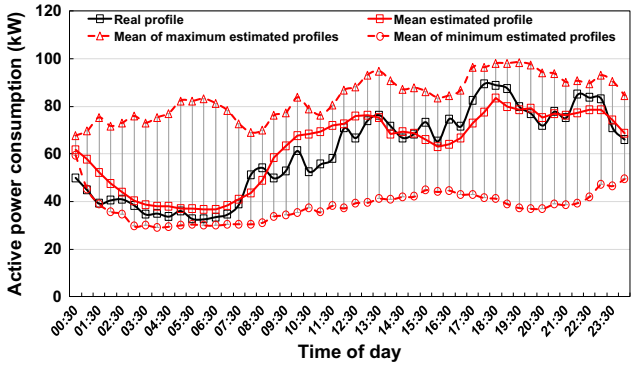

Fig. 6. Actual and estimated load profiles of a weekday.

Manhattan (city block) distance functions results in the smallest error distributions.

Canberra distance function, as compared to Manhattan, produces more accurate load estimates. Fig. 8 indicates that 75% of the errors were less than 10% (or 10 kW) when the load was estimated using Canberra distance. Fig. 8 also shows that the application of Canberra distance function resulted in large values of the maximum MAPE. These extreme values of the MAPE were observed when short segmented cluster centres were applied to estimate the load of aggregated smart meters. The application of long segmented cluster centres results in smaller values (and therefore smaller distributions) of the MAPE. In Fig. 8, the dark shaded boxes represent the distribution of estimation errors between the first quartile and median of the MAPE (and RMSE). Load estimation errors between the median and the third quartile of MAPE (and RMSE) are represented by the light-shaded boxes. Error bars represent the minimum and maximum values of the

Please cite this article in press as: Al-Wakeel A et al. *k*-means based load estimation of domestic smart meter measurements. Appl Energy (2016), http://dx.doi.org/10.1016/j.apenergy.2016.06.046

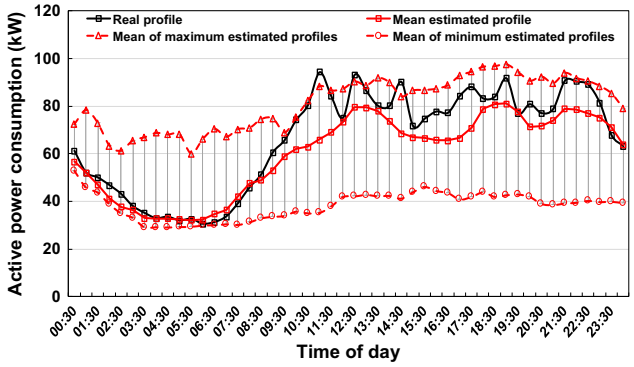

Fig. 7. Actual and estimated load profiles of a weekend.

estimation errors. The horizontal line that splits the dark-shaded and light-shaded boxes is the median of the MAPE (and the RMSE) [35].

5.2. Impact of the daily and segmented cluster centres

Simulation results show that the application of segmented cluster centres for load estimation results in more accurate estimates than the application of daily cluster centres. Fig. 9 compares the distribution of the MAPE (and RMSE) when segmented and daily cluster centres were applied to estimate the missing measurements. Fig. 9 shows that regardless of the duration of missing measurements, the application of segmented cluster centres produces significantly less estimation errors than daily cluster centres. However, as compared to the daily centres, the application of segmented cluster centres results in higher maximum values of estimation errors. The application of short (2–4 h) segmentation time windows results in these errors. The errors can be ignored by using longer cluster centres to estimate the missing measurements.

5.3. Impact of the segmentation time window

The time window of segmented cluster centres that provided the smallest estimation errors was between 16 and 24 h. Fig. 10 shows a box-whisker plot of the distribution of MAPE (and RMSE) of the estimated measurements. The distribution of estimation errors was uniform around the median for all segmentation time windows. The smallest values of minimum and maximum errors (error bars) were observed in the 16–24 hours' range of segmentation time windows. Fig. 10 illustrates that the maximum load estimation errors occurred at short segmentation time windows.

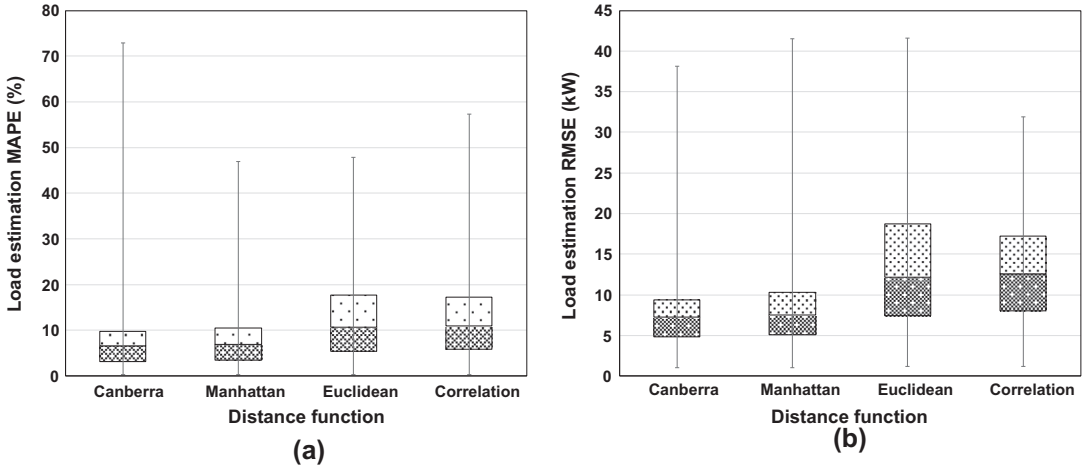

Fig. 8. Distribution of the load estimation errors for different distance functions, (a) MAPE and (b) RMSE.

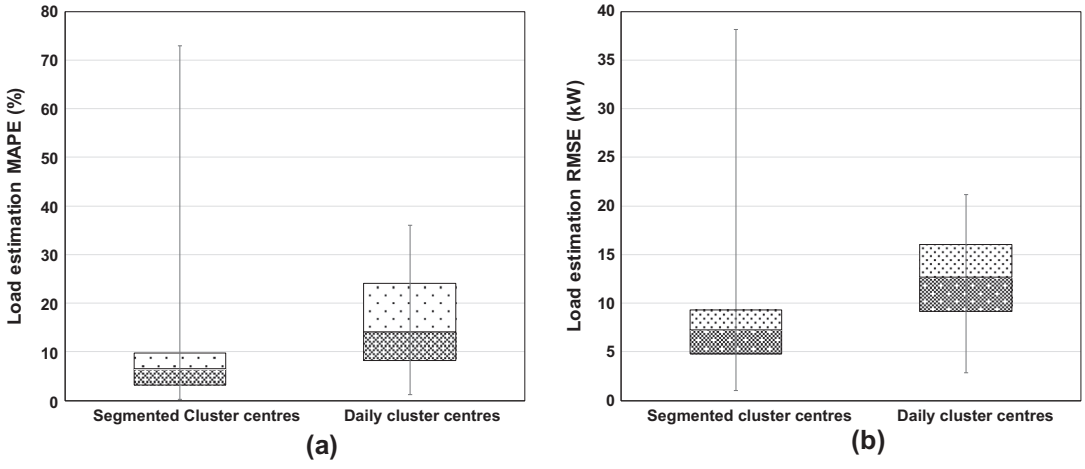

Fig. 9. Distribution of load estimation errors for different durations of cluster centres, (a) MAPE and (b) RMSE.

Please cite this article in press as: Al-Wakeel A et al. *k*-means based load estimation of domestic smart meter measurements. Appl Energy (2016), http://dx.doi.org/10.1016/j.apenergy.2016.06.046

5.4. Impact of the duration of measurement loss

Simulation results indicate that the load estimator was capable of providing load estimates for different durations of measurement loss. Fig. 11 shows a box-whisker plot representation of the distribution of the MAPE (and RMSE) up to 24 h of measurement loss.

The average daily power consumption of the aggregated smart meter measurements during the test period was approximately 61 kW for working days and 65 kW for the weekends. Fig. 11 shows that when the maximum values of the MAPE are neglected, up to 9 h of measurements were estimated with a MAPE that was equal to 5%. This value of MAPE corresponds to approximately 10% of the average daily consumption during the test period.

5.5. Performance of the load estimation algorithm

The performance of the load estimation algorithm was compared with other available methods which have been used for load estimation. These methods include the NAÏVE estimator, Linear Auto-Regressive eXogenous (ARX) model, Non-linear Neural Network (NN) model, Linear regressive model using the Least Mean Squares (LMS) algorithm, and Non-linear Auto-Regressive eXogenous (NARX) model that were used in [36] to estimate the demand at MV busbars.

The load estimation algorithm used one week of aggregated smart meter measurements to train the algorithm and one further week to test and validate the algorithm. Smart meter measurements were collected from the Irish smart metering trials as presented in Section 3. However, the load estimation methods reported in [36] utilised 24 months of weather data and aggregated smart measurements that were collected from a test distribution network in Denmark. The Danish data was split into 12-month model training data and 12-month model validation data. Because the implementation and the dataset used for each method were different, the comparison was mainly based on 24-h ahead load estimation errors reported in [36] and obtained from the simulation results.

The Mean MAPE reported in the literature varied among 8% (NARX), 9% (ARX), 10% (NN), 11% (LMS) and 12% (NAÏVE). The errors of the developed load estimation method are shown in Figs. 8–11 which shows its good performance. Fig. 11 shows that the MAPE of the estimated load was less than 2% up to 4 h of missing measurements.

Simulation results reveal that the increase in clustering errors (between the load profiles and their corresponding cluster centres) decreases the load estimation accuracy. Fig. 12 shows an illustration of load estimation MAPE versus the clustering errors (there are 24 curves for each sub-figure which represent the load estimation MAPE from 1 h up to 24 h of missing measurements). The solid red error profile is the mean of the maximum MAPE (Fig. 12a) and the mean of the mean MAPE (Fig. 12b) of the estimated load. The dashed red profiles represent the maximum and minimum values of the maximum and mean MAPEs of load estimates.

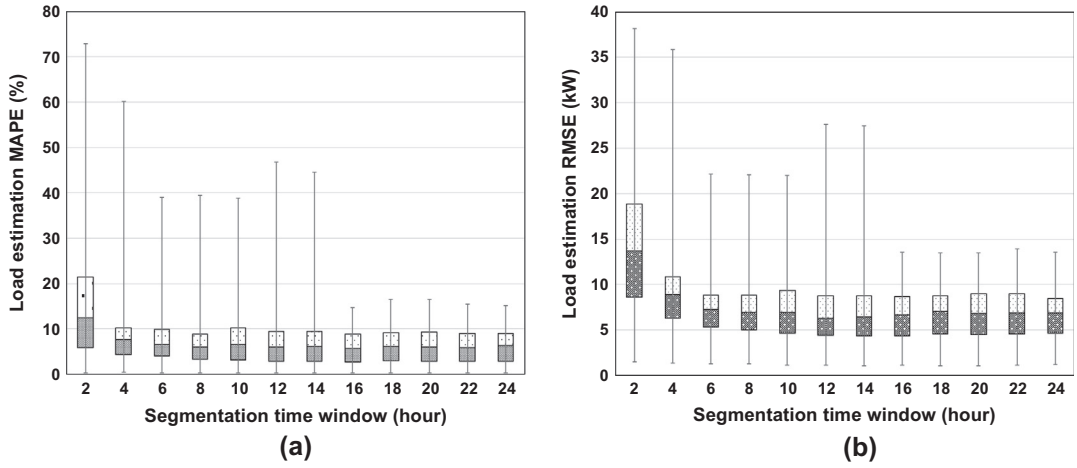

Fig. 10. Distribution of load estimation errors for different segmentation time windows, (a) MAPE and (b) RMSE.

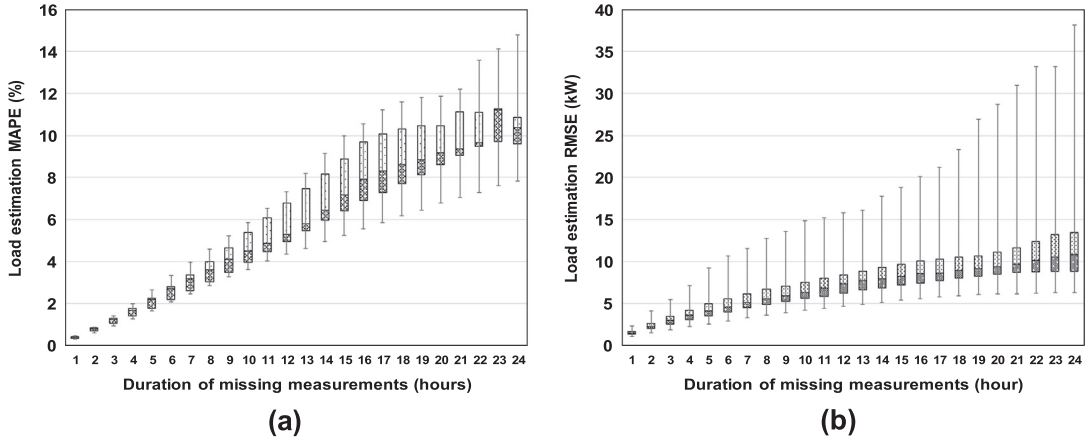

Fig. 11. Distribution of load estimation errors for different durations of measurement loss, (a) MAPE and (b) RMSE.

Please cite this article in press as: Al-Wakeel A et al. *k*-means based load estimation of domestic smart meter measurements. Appl Energy (2016), http://dx.doi.org/10.1016/j.apenergy.2016.06.046

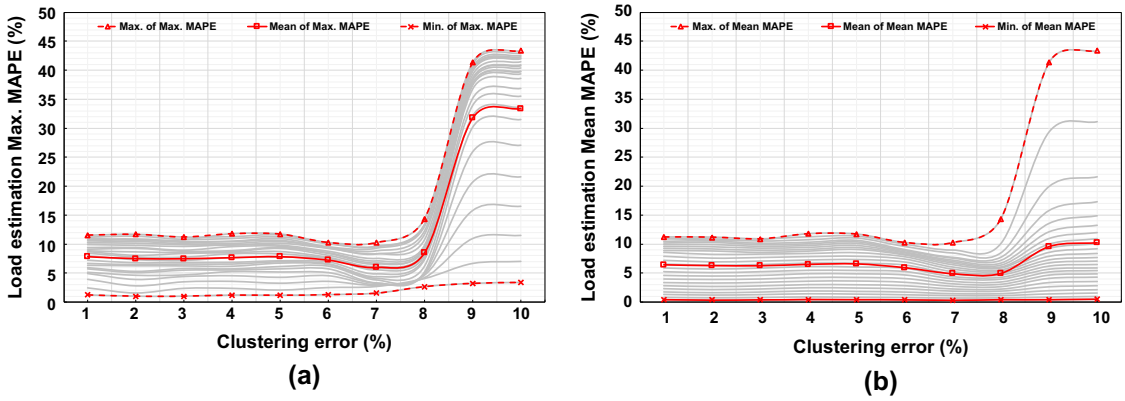

Fig. 12. Load estimation errors vs. clustering errors (a) maximum MAPE, (b) mean MAPE.

6. Conclusions

A load estimation algorithm based on *k*-means cluster analysis method was developed. Detailed analysis of the simulation results showed that Canberra distance function was capable of providing accurate load estimates as compared to other functions. The application of segmented cluster centres proved to be more effective than daily centres. Segmented cluster centres with a length in the range of 16–24 h resulted in higher accuracy load estimates than other lengths of segmentation.

Acknowledgements

The authors gratefully acknowledge the EPSRC Increasing the Observability of Electrical Distribution Systems using Smart Meters (IOSM) Project (Grant No. EP/J00944X/1), the P2P-SmarTest Programme (H2020-646469) through the European Commission HORIZON 2020 grant, and the UK–China NSFC/EPSRC OPEN Project (Grant No. EP/K006274/1 and 51261130473) for the partial support of this work. Information about the data that underpins the results presented in this paper, including how to access those data, can be found in Cardiff University's data catalogue at http://dx.doi.org/10.17035/d.2016.0009225471.

References

[1] Coelho VN, Coelho IM, Coelho BN, Reis AJR, Enayatifar R, Souza MJF, et al. A self-adaptive evolutionary fuzzy model for load forecasting problems on smart grid environment. Appl Energy 2016;169:567–84.
[2] DCC. Smart meter key infrastructure, infrastructure key infrastructure and DCC key infrastructure; 2015 [Online Available: Smart Meter Key Infrastructure, Infrastructure Key Infrastructure and DCC Key Infrastructure, accessed: 18-Feb-2016].
[3] DECC. Smart meters, smart data, smart growthAvailable from: <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/397291/2903086_DECC_cad_leaflet.pdf>2015 [accessed: 27-Aug-2015].
[4] Jenkins N, Long C, Wu J. An overview of the smart grid in Great Britain. Engineering 2015;1(4):413–21.
[5] DCC. Building a smart metering network for Great BritainAvailable from: <https://www.smartdcc.co.uk/media/338770/15574_building_a_smart_metering_network_v3.pdf>2015 [accessed: 28-Feb-2016].
[6] Carvallo A, Cooper J. The advanced smart grid edge power driving sustainability. Artech House; 2015.
[7] Danahy Jack. That smart grid data surge we mentioned earlier? You can't ignore itAvailable from: <http://www.smartgridnews.com/story/smart-grid-data-surge-we-mentioned-earlier-you-can-t-ignore-it/2009-11-03>2009 [accessed: 21-Feb-2016].
[8] Bristol Smart Energy City. Bristol smart energy city collaboration Appendix B: technical – data and ITAvailable from: <https://bristol-smart-energy.cse.org.uk/wiki/B:_Technical_-_Data_and_IT#T1:_Application_scales_for_datasets>2015 [accessed: 21-Feb-2016].
[9] Ofcom. Infrastructure report 2014. Ofcom's second full analysis of the UK's communications infrastructure; 2014 <http://stakeholders.ofcom.org.uk/binaries/research/infrastructure/2014/infrastructure-14.pdf>.
[10] Tafazolli R. Smart metering system for the UK technologies reviewAvailable from: <https://m2m.telefonica.com/system/files_force/SM_Report_07-06-2013 2_9.pdf?download=1>2013 [accessed: 27-Aug-2015].
[11] Commission for Energy Regulation. Electricity smart metering technology trials findings report. Information Paper CER11080b; 2011 <https://www.ucd.ie/t4cms/ElectricitySmartMeteringTechnologyTrialsFindingsReport.pdf> [accessed: 01-Sep-2015].
[12] McLoughlin F, Duffy A, Conlon M. A clustering approach to domestic electricity load profile characterisation using smart metering data. Appl Energy 2015;141:190–9.
[13] Singh R, Pal BC, Jabr RA. Statistical representation of distribution system loads using Gaussian mixture model. IEEE Trans Power Syst 2010;25(1):29–37.
[14] Tawalbeh NIA. Daily load profile and monthly power peaks evaluation of the urban substation of the capital of Jordan Amman. Int J Electr Power Energy Syst 2012;37(1):95–102.
[15] Soares LJ, Medeiros MC. Modeling and forecasting short-term electricity load: a comparison of methods with an application to Brazilian data. Int J Forecast 2008;24(4):630–44.
[16] Zhou K, Fu C, Yang S. Big data driven smart energy management: from big data to big insights. Renew Sustain Energy Rev 2016;56:215–25.
[17] Raza MQ, Khosravi A. A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings. Renew Sustain Energy Rev 2015;50:1352–72.
[18] Hand DJ, Mannila H, Smyth P. Principles of data mining. Massachusets Institute of Technology; 2001.
[19] Zaki MJ, Meira Jr W. Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press; 2014.
[20] Räsänen T, Voukantsis D, Niska H, Karatzas K, Kolehmainen M. Data-based method for creating electricity use load profiles using large amount of customer-specific hourly measured electricity use data. Appl Energy 2010;87(11):3538–45.
[21] Räsänen T, Ruuskanen J, Kolehmainen M. Reducing energy consumption by using self-organizing maps to create more personalized electricity use information. Appl Energy 2008;85(9):830–40.
[22] Willis H, Schauer A, Northcote-green JD, Vismor T. Forecasting distribution system loads using curve shape clustering. IEEE Trans Power Apparat Syst 1983;PAS-102(4):893–901.
[23] Panapakidis IP. Clustering based day-ahead and hour-ahead bus load forecasting models. Int J Electr Power Energy Syst 2016;80:171–8.
[24] Bandyopadhyay S, Saha S. Unsupervised classification: similarity measures, classical and metaheuristic approaches, and applications. Springer-Verlag; 2013.
[25] Gan G, Ma C, Wu J. Data clustering: theory, algorithms, and applications. SIAM; 2007.
[26] Mena R, Hennebel M, Li Y-F, Zio E. Self-adaptable hierarchical clustering analysis and differential evolution for optimal integration of renewable distributed generation. Appl Energy 2014;133:388–402.
[27] Han J, Pei J, Kamber M. Data mining: concepts and techniques. Elsevier; 2011.
[28] Rhodes JD, Cole WJ, Upshaw CR, Edgar TF, Webber ME. Clustering analysis of residential electricity demand profiles. Appl Energy 2014;135:461–71.
[29] Irish Social Science Data Archive. Irish smart metering measurements <http://www.ucd.ie/issda/data/commissionforenergyregulationcer/> [accessed: 10-Mar-2016].
[30] Zivot E, Wang J. Modeling financial time series with S-PLUS. Springer-Verlag; 2006.
[31] de Hoon MJL, Imoto S, Nolan J, Miyano S. Open source clustering software. Bioinformatics 2004;20(9):1453–4.

[32] Everitt BS, Landau S, Leese M, Stahl D. Cluster analysis. John Wiley & Sons Ltd.; 2011.
[33] Jain K, Anil, Dubes C, Richard. Algorithms for clustering data. Prentice-Hall, Inc.; 1988.
[34] Tukey JW. Exploratory data analysis. Addison-Wesley Publishing Company; 1977.
[35] Der G, Everitt BS. Basic statistics using SAS® enterprise guide: a primer. SAS Institute; 2007.
[36] Hayes BP, Prodanovic M. State forecasting and operational planning for distribution network energy management systems. IEEE Trans Smart Grid 2016;7(2):1002–11.

Please cite this article in press as: Al-Wakeel A et al. *k*-means based load estimation of domestic smart meter measurements. Appl Energy (2016), http://dx.doi.org/10.1016/j.apenergy.2016.06.046